CLAIMS

1. A method of operating on a text comprising a plurality of text units, each comprising one or more strings, the method being characterised by:

forming a structure for each of at least some of said strings, in which structure a string is associated with each pair of text units in which the string occurs;

for each pair of text units summing the number of occurrences of each other text unit in the same structure or structures so as to form an individual score for each pair of text units; and

processing said individual scores for each pair of text units in order to form a final score for each pair of text units to determine how many times any string is shared between each pair of text units and other text units.

2. A method of operating on a text as claimed in claim1, which includes the further step of ranking the text units on the basis of said individual scores.

3. A method of operating on a text as claimed in claim 1, wherein said text units are sentences, said

strings are words forming said sentences, and the method comprises the additional steps of removing stop-words, stemming each remaining word and indexing the sentences prior to carrying out said summing step, and wherein said structures are stem-index records each comprising a stemmed word and one or more indexes corresponding to sentences in which said stemmed word occurs.

4. A method of operating on a text as claimed in claim 1, wherein said text is associated with a word text comprising words, each word being associated with one or more subject codes representing subjects with which said word is associated, and wherein said strings are subject codes associated with said words.

5. A method of operating on a text as claimed in claim 4, which comprises the further step of keeping a record of the word spelling associated with each occurrence of a subject code in a text unit, and wherein during said summing step occurrences of the same subject code in a pair of text units are disregarded if the same word spelling is associated with said same subject code in said pair of text units.

6. A method of operating on a text as claimed in claim 5, wherein said step of disregarding occurrences of subject codes is not carried out for subject codes which relate to only a single word spelling in the word text.

7. A method of operating on a text as claimed in claim 1, wherein said processing step includes calculating a level for each text unit, in addition to said final score, and wherein said level indicates the value of the highest of said individual scores in relation to a threshold value.

8. A storage medium containing a program for controlling a programmable data processor (70) to perform a method as claimed in claim 1.

9. A system for ranking text units in a text, the system comprising a data processor (70) programmed to perform the steps of the method of claim 1.